Attacking the Power-Wall by Using Near-Threshold Cores

Liang Wang
Department of Computer Science
Univiersity of Virginia
liang@cs.virginia.edu

ABSTRACT

Due to limited scaling of supply voltage, power density is expected to grow in future technology nodes. This increasing power density limits the number of simultaneously switching transistors for future chips, leaving others inactive, a phenomenon referred to as "dark silicon". This project investigates the idea of improving throughput with higher utilization of die area by exploiting near-threshold voltage operation instead of dark silicon. To test the hypothesis, an analytical model is developed to quantify the performance of systems with near-threshold cores. Results confirm improvements in both throughput and utilization of the chip. However, these improvements are getting less and less regarding the impact of process variations.

1. INTRODUCTION

Due to limited scaling of supply voltage, conventional multicore scaling is getting more and more challenging these days. This scaling trend in turn is due to constraints on threshold voltage. Threshold voltage scales down more slowly in current and future technology nodes to keep leakage power under control. In order to achieve fast switching speed, it is generally necessary to keep transistors operating at a considerably higher voltage than their threshold voltage. Therefore, the dwindling scaling on threshold voltage leads to a slower pace of supply voltage scaling. Since the switching power of a single transistor changes quadratically to the supply voltage, the scaling of the switching power would violates Dennard Scaling in future technology nodes. On the other hand, design constraints, such as cooling cost and on-chip power delivery limitations, hinder further increase in the thermal design power (TDP) of a chip. As Moore's Law continues to double transistor density across technology nodes, the total power consumption will soon exceed TDP, with all available transistors actively switching at their nominal speed. As a result, future chips would only support a small fraction of transistors, leaving others inactive, a phenomenon referred to as "dark silicon" in [13].

This project investigates the idea of improving throughput with higher utilization of die area by exploiting near-threshold voltage operation instead of dark silicon. By lowering voltage, the dynamic power of a single transistor is reduced dramatically. In consequence, more cores can be activated at the same time, and the chip utilization is improved as well. More parallelism has the potential to compensate for frequency loss by near-threshold operation. Since operations of a single core are slower than its nominal speed, a system composed by those cores is called "dim silicon." In this project, we hypothesize that near-threshold voltage operation improves the performance for throughput computing over conventional super-threshold system organizations, which suffer from "dark silicon" in near-future technology nodes. To test the hypothesis, an analytical model is developed to quantify the performance of the proposed organization. It extends Amdahl's Law with frequency/voltage scaling in the near-threshold region. Results confirm improvements in both throughput and utilization of the chip.

The following paper is organized as follows: the second section describes the background and related works, the third section dives into more details about analytical model and evaluation methodology, the fourth section presents various architectural analyses and results, and the final section summarizes the work.

2. RELATED WORK

2.1 Scaling Trends

The power issue in future technology scaling has been recognized as one of the most important design constraints by architecture designers [13, 9]. Esmaeilzadeh et al. did a comprehensive design space exploration on future technology scaling with an analytical performance model [5]. While primarily focusing on single-core performance, they did not consider lowering supply voltage to have dim silicon on a chip. Without such consideration, they claimed future chips would inevitably suffer from a large portion of dark silicon at deeply scaled technology nodes. In [1], Borkar and Chien suggested the approach of near-threshold computing by aggressive supply-voltage scaling, and indicated its potential benefits for aggregate throughput improvement. In contrast, our proposed project studies near-threshold computing quantitatively in more detail with the help of an analytical model. In [6], Huang et al. have done a design space exploration on future technology nodes with analytical models. They recommended dim silicon and briefly mentioned the possibility of near-threshold operation. They also explored

relaxing constraints for area and TDP along with novel cooling solutions to maintain ideal 2X throughput growth. Although using similar methodology and scaling trends, our proposed project evaluates in detail the potential benefit of improving aggregate throughput by near-threshold computing.

2.2 Near-Threshold Operation

Near-threshold computing(NTC) operates circuits at a lower voltage close to their threshold voltage. It provides the benefit of low dynamic power consumption and higher energy efficiency, as shown in [4]. However, there are a couple of issues associated with NTC. Firstly, the switching speed of a transistor slows down due to small over-drive voltage (supply voltage minus threshold voltage). Secondly, a single transistor is more sensitive to threshold variation when supply voltage is getting close to its threshold, leading to a significant increase in performance variation. Finally, variations in process, temperature and voltage makes the functionality of circuits more vulnerable, especially for SRAM. In addition to pointing out those issues, Dreslinsk et al. had surveyed and summarized various techniques to accommodate those issues in [4]. They had also suggested NTC integration in ultra energy-efficient servers with high throughput. Our project studies NTC in detail with quantitative results to show the NTC effectiveness in high throughput computing. Even suffering from those drawbacks, near-threshold computing shows a potential to deliver high performance with throughput computing. In [7], Krimer et al. had demonstrated a near-threshold stream processor with SIMD architecture for energy-constrained throughput computing. Instead of energy-constrained throughput, our proposed project focuses on power-constrained throughput. We do not target at one specific system architecter, but more generally, our proposed project studies near-threshold computing more broadly with various types of system architectures in different technology nodes. It quantifies the potential of near-threshold computing as well as its limitations in context of power-constrained scaling.

3. CIRCUIT SIMULATION

In order to get more accurate scaling relationship between supply voltage and frequency, a bunch of circuit simulations are employed in this project. Two circuits are simulated, a single inverter and a ripple carry adder. Their schematics are shown in Figure 1. The width of the adder is varied from 4bits up to 32bits, stepping by the power of 2. All circuits are simulated with Spectre driven by Ocean scripts.

4. METHODOLOGY

We use aggregate throughput under TDP constraints as the primary performance metric. Instead of running extensive architectural simulation, we propose analytical models based on Amdahl's Law and use them to evaluate system performance with various organizations. Systems are modeled as symmetric multi-core organizations composed by either simpler in-order cores or more complex out-of-order cores. Two core designs at 45nm are picked up from McPAT[8] as baseline cores, a Niagara2-like in-order core (IO) and a Penryn-like out-of-order core (O3). The characteristics of a single core are obtained from McPAT and summarized in Table 1.

	Frequency (GHz)	Dynamic Power (W)	Leakage Power (W)	Area (mm^2)
IO	4.20	6.14	1.06	7.65
O3	3.70	19.83	5.34	26.48

Table 1: Baseline cores at 45nm with high performance process.

Systems are studied across technology nodes ranging from 45nm down through 16nm. For process related scalings, a modified version of Predictive Technology Model (PTM)[3] is used, which provides a more realistic perspective on original PTM release [10]. For each technology node, there are two process variants as a high performance High-K metal gate silicon process and a low power process, respectively. The nominal supply voltages and threshold voltages are characterized in Table 2

		$45nm$	$32nm$	$22nm$	$16nm$
High Perf.	V_{nom} (V)	1.0	0.9	0.8	0.7
	V_t (mV)	424.25	466	508.16	504.9
Low Power	V_{nom} (V)	1.1	1.0	0.95	0.9
	V_t (mV)	622.61	647	707.3	710.32

Table 2: Nominal supply voltage and threshold voltage for each PTM technology library

The model proposed in this project extends Amdahl's Law with the following three extensions: *near-threshold frequency scaling*, *power model*, and *performance model*. They are described in sequence in the rest of this section.

4.1 Frequency Scaling

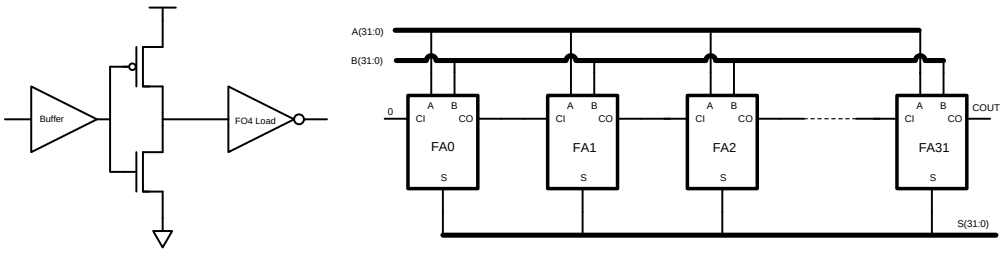
A couple of formula-based first-order approaches have been proposed to model the frequency scaling when sweeping supply voltage, such as α-power law in [12] and unified transistor model in [11]. However, α-power law is only valid for super-threshold scaling, where supply voltage is much higher than threshold voltage; the unified transistor model requires process-specific parameters and it is less accurate than real simulation results. Therefore, this project exploits real circuit simulation, and derive analytic models by fitting results from the simulations.

In order to model the core frequency, I simulate a single inverter and a 32-bits adder. The simulation results show a similar scaling trend between this two circuits. Hence, the following analysis will be based on inverter if not mentioned.When it comes to the variation, the inverter circuit is too simple to match the complexity of a processor core. In this case, simulation results on adder will be used.

4.2 Power Modeling

Core power consumption comes from two major sources, dynamic power due to transistor switching and static power due to leakage. Therefore, per-core power is given by

$$P_{\text{total}} = P_{\text{dynamic}} + P_{\text{leakage}} \quad (1)$$



(a) Inverter (b) Adder

Figure 1: Circuits schematics

Generally speaking, dynamic power is given by

$$P_{\text{dynamic}} = C_{eff} \cdot V_{dd}^2 \cdot f \tag{2}$$

where C_{eff} is the effective capacitance, V_{dd} is the supply voltage, and f is the core running frequency. We assume a constant effective capacitance during VFS in this project. Therefore, dynamic power changes quadratically to supply voltage and linearly to frequency. According to [2], the static power of a system is given by

$$P_{\text{leakage}} = V_{dd} \cdot N \cdot k_{design} \cdot \hat{I}_{leak} \tag{3}$$

where V_{dd} is the supply voltage, N is the number of transistors, k_{design} is a device-specific constant for a given transistor, and $\hat{I}_{leak}$ is the normalized per-transistor leakage current. We assume the same leakage current for all transistors in this project. Therefore, $\hat{I}_{leak}$ is identical to leakage current of a single transistor. According to [11], the leakage current of a single transistor is given by

$$I_{leak} = I_0 \cdot 10^{\frac{V_{gs} - V_t + \eta V_{ds}}{S}} \cdot \left(1 - e^{-\frac{V_{ds}}{V_{th}}}\right),$$
$$S = n \cdot V_{th} \cdot \ln 10$$

where V_t is the threshold voltage, η is the drain-induced barrier lowering factor (DIBL), V_{th} is the thermal voltage, and n is the process dependent constant. For simplicity, we do not consider body-bias effect in this model, so V_t is only dependent on specific manufacturing process. The thermal voltage at room temperature is around $28mV$, which is far less than the supply voltage of interest in this project, therefore $e^{-V_{ds}/V_t} \approx 0$. Because the transistor is at its static state when considering the static leakage power, V_{gs} and V_{ds} is roughly proportional to the supply voltage. As a result, the above equation can be deducted to

$$\hat{I}_{leak} \propto 10^{\frac{V_{dd}}{\hat{S}_{leak}}} \tag{4}$$

where $\hat{S}_{leak}$ is the aggregate scaling coefficient derived from fitting to the simulated results.

4.3 Performance Modeling

As for the aggregate throughput performance, we model it with Amdahl's Law as shown in Equation 5

$$\text{Speedup} = \frac{1}{\frac{1-\rho}{S_{\text{serial}}} + \frac{\rho}{n \cdot S_{\text{parallel}}}} \tag{5}$$

where ρ is the parallel ratio of the studied workload, S_{serial} is the serial part speedup over a baseline core, S_{parallel} is the per-core speedup when the workload is run in parallel, n is the number of active cores running in parallel. For S_{serial}, only one core is utilized. For simplicity, we assume the core runs at the highest frequency to achieve the best single core performance, denoted as pf_c. For parallel part, both n and S_{parallel} are determined by supply voltage. By applying VFS, per-core frequency scales with supplying voltage, therefore per-core performance is a function of supply voltage as $pf(v)$. Additionally, per-core power changes with supply voltage which is denoted as $p(v)$. The number of active cores is restricted by the budgets of power (P) and area (A), which is given by

$$n(v) = \min(\frac{P}{p(v)}, \frac{A}{a}) \tag{6}$$

As a result, Equation 5 can be rewritten as

$$\text{Speedup} = 1 \Big/ \frac{1-\rho}{\frac{pf_c}{pf_0}} + \frac{\rho}{n(v) \cdot \frac{pf(v)}{pf_0}} \tag{7}$$

where pf_0 is the performance of a single baseline core.

5. ANALYSES AND RESULTS

5.1 Voltage-to-Frequency Scaling

Two process variants are plotted for voltage-to-frequency scaling for each technology nodes. The results for 45nm is shown in 2.

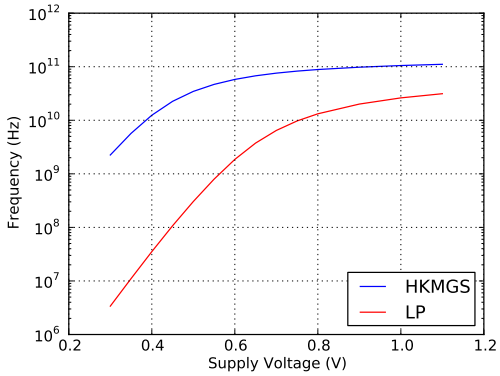


Figure 2: Frequency-to-voltage scaling, inverter at 45nm

Since low-power process has a higher threshold voltage than high performance process, circuits with low power process

reach the near-threshold operating region earlier than high performance process, when scaling down supply voltage from $1.1V$ down to $300mV$. As a result, circuits with low power process has a much larger frequency loss than high performance process with the same change in supply voltage. For example, when scale down supply voltage from $800mV$ to $400mV$, low power process suffers frequency loss by 400x while high performance process only decreases by around 8x.

5.2 Performance Improvement via VFS

Each core in the symmetric multi-core system is capable of scaling its supply voltage from 1.1 volts down to as low as 0.3 volts. Area and power are the two primary design constraints for such a system. Area determines the number of cores available on a chip, while power confines the number of active running cores.

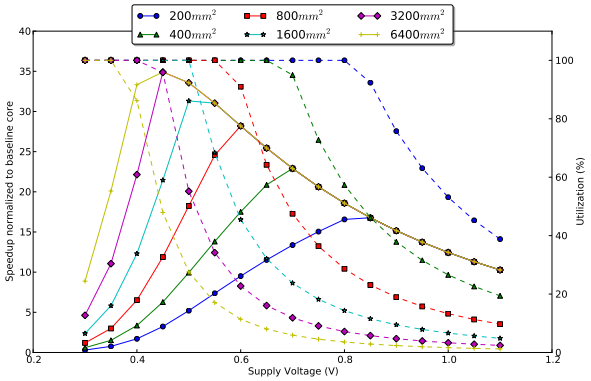


Figure 3: Performance scaling for in-order cores (IO) under 45nm HKMGS technology, the system has the TDP of 100W

Firstly, the power budget of the system holds. In Figure 3 corresponds to systems with the same power budget but various area budgets. With a given power budget, speedup curves are identical to overlap with each other in the upper range of voltages. This is because in such a voltage region, per-core power is relatively large so that the number of active cores is mainly constrained by the power budget. Those speedup curves eventually reach their peak, and further decrease in supply voltage hurts overall speedup. That is because the system comes to the point where per-core power is small enough to have all on-chip cores activated at the same time. With further decrease in supply voltage, the number of active cores does not increase any more, but per-core performance is reduced due to the lower supply voltage. Since the number of total on-chip cores is determined by the chip area, different area budgets lead to different voltage of speedup peak. Larger area permits a larger number of active cores at peak performance point. Thus, a system with larger area achieves better speedup. The only exception is when the system area increases from $3200mm^2$ to $6400mm^2$. In this case, the system does not gain any increase in its best achievable speedup with a larger area. It conveys the fact that the increase in number of cores fails to overcome the frequency drawback of lower voltage. Since increased area does not introduce any performance gains, we call it "power limited."

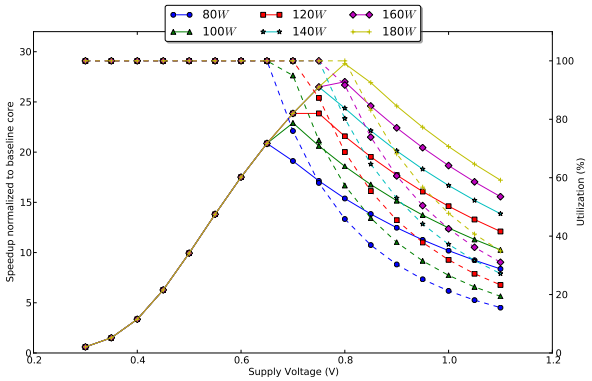


Figure 4: Performance scaling for in-order cores (IO) under 45nm HKMGS technology, the system has the fixed area of 400mm^2

Secondly, the area budget of the system holds. In Figure 4, each series has the same area budget but various power budgets. Speedup curves have demonstrated similar escalate-decline trends. Those speedup curves converge with each other from their peak point towards the lowest supply voltage. Speedup curves converge at the upper section of voltage scaling spectrum. These convergences imply that the speedup of a system running at a higher voltage is more sensitive to its power budget, and the speedup of a system running at a lower voltage is more sensitive to its area budget. Moreover, in Figure 3, there is no such "area limited" scenarios. A system with higher power budget always ends up with higher optimal performance speedup. A system scales down its supply voltage to the point where all on-chip cores are active to reach its optimal speedup. Although the number of active cores do not increase at that point with a higher power budget, the surplus power can be used to increase per-core frequency, and the overall speedup eventually benefits from per-core performance improvement.

5.3 Across Technology Studies

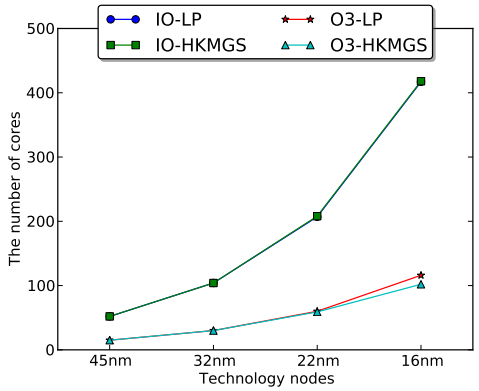


Figure 6: Optimal number of cores for each technology nodes, the system has area of 400mm^2 and power of 100W.

In the last section, a system exhibits its best performance with supply voltage in between the maximum $1.1V$ and min-

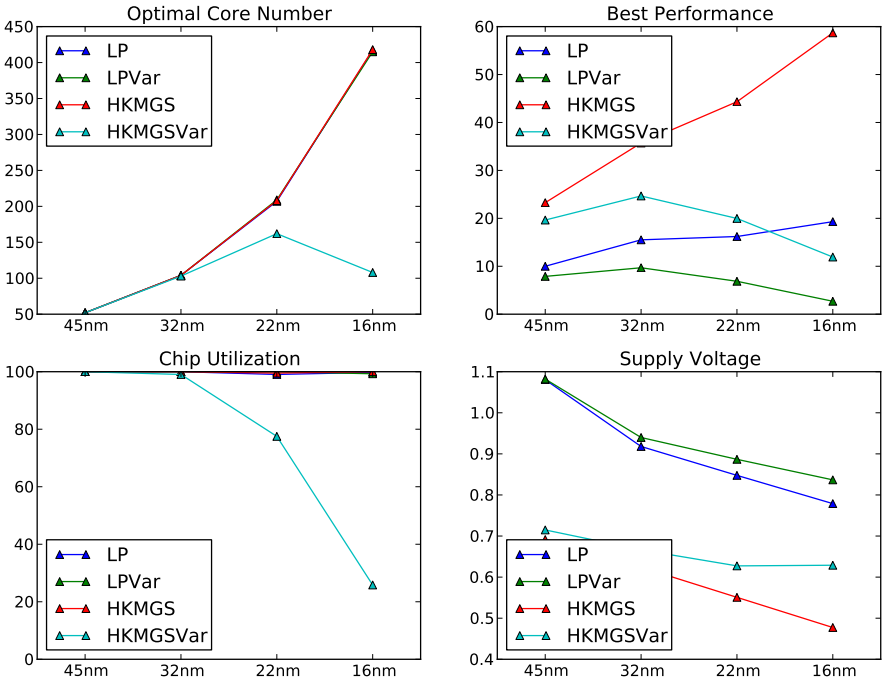


Figure 5: Impact of process variations, the system has area of $400mm^2$ and power of $100W$.Series names ending with 'var' indicate the variation-aware results.

imum $300mV$. In this section, I will focus on the optimal supply voltage where a system achieves its best performance. The optimal supply voltage is searched by comparing every system configurations with different number of cores activated at the same time. For a given number of active cores, the supply voltage is optimized to its maximum as long as the system stays under the total power budget. The number of active cores with the best performance is plotted in Figure 6. The plot shows an 2x increase in core numbers for the most combinations of core types and process types. In this case, the optimal core number is actually the maximum number of cores under the area constraint, which, in other words, the system is fully utilized when achieving its best performance and completely eliminates dark silicon across technology nodes.

Besides the utilization of the chip, I have also compared the speedup of these system configurations, which is plotted in Figure 7. In this plot, high performance process demonstrate higher aggregate speedup comparing to its low power counterpart. The speedup gap between the high performance process and the low power process increases from around 2x at 45nm to around 3x at 16nm. Such scaling trends suggest high performance process a better candidate for future technology nodes if aggregate performance of the system is the primary metric for chip designers.

5.4 Variation

As stated in Section 2, process variations play an important role with near-threshold operations in both performance and functionality. In this paper, we only consider the variation impact on performance, leaving the functionality for future work. Since the system studied in this paper is on-

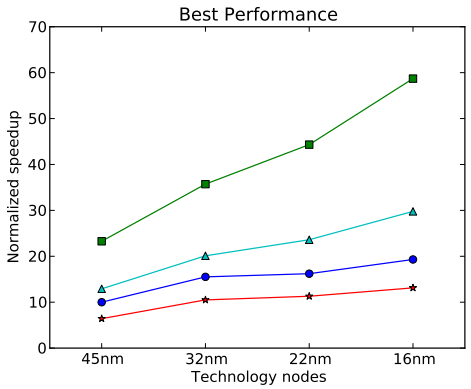


Figure 7: The best speedup for each technology nodes, the system has area of $400mm^2$ and power of $100W$.

chip multi-core, the global variation is consistent among all cores. However, cores may suffer from local variations that differs core to core. A single inverter becomes too simple to model the complexity and various local variations of a general-purpose processor core. As a result, a 32bits adder is chosen for variation-aware studies in this section.

The impact on frequency at 45nm is plotted in Figure 8. Frequencies drop down by up to 5x for high performance process, and 10x for low power process. Such frequencies slow down is more severe in future technology nodes, for example, 10x and 100x slow down for 16nm, high performance and lower power process, respectively. The increasing fre-

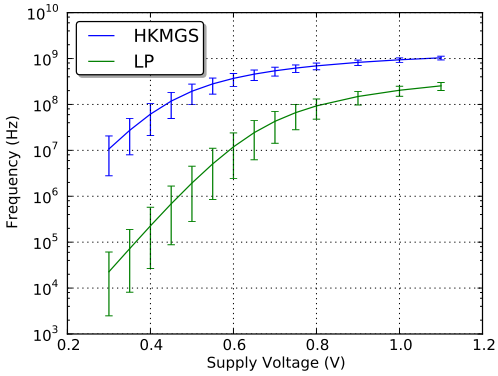


Figure 8: Variation aware circuit simulation on 32bits adder at 45nm

quency drop-down has critical performance impact on the whole system, as plotted in 5.

As for the number of active cores, only the high performance process suffers from variation, experiencing the chip utilization as low as 20% at 16nm. However, when it comes to the performance metric, both high performance process and low power process come up with performance slow down starting from 32nm. The aggregate throughput is dictated by the slowest core among all cores in the system, which tends to be increasingly slowed in future technology node. While, on the other hand, the power consumption roughly remains the same, it limits how low the supply voltage can go for future technology nodes. Putting it as a whole, the overall throughput is significantly limited with future technology nodes.

6. CONCLUSION

In terms of overall performance, the system with simple in-order cores is better than the system with complex out-of-order cores, and high performance process is actually better than low power process. Lowering the supply voltage greatly reduces the percentage of dark silicon on chip, delivering higher aggregate throughput than dark system. However, the process variation hurts the performance improvement severely. In some cases, it is even slower than the dark system. To wrap up into a conclusion, cores operating at near-threshold voltage is effective in improving performance and reducing dark silicon, only when a great effort has be taken on controlling the negative effect of process variation.

7. ACKNOWLEDGEMENT

I would like to thank Prof. Calhoun for his advices, suggestions and encouragement for this project, as well as the course all through the semester. I would also thank TAs, Kyle Craig and Yousef Shakhsheer, for their help on setting up the circuit simulation infrastructure. I would thank Yanqing Zhang for his help on circuit synthesis with RC Compiler, and Jim Boley for his help on Monte-Carlo simulations.

8. REFERENCES

[1] S. Borkar and A. A. Chien. The Future of Microprocessors. *Communication of the ACM*, 54(5):67–77, May 2011.

[2] J. A. Butts and G. S. Sohi. A Static Power Model for Architects. In *International Symposium on Microarchitecture*, MICRO '00, pages 191–201, 2000.

[3] B. H. Calhoun, S. Khanna, R. Mann, and J. Wang. Sub-threshold Circuit Design with Shrinking CMOS Devices. In *International Symposium on Circuits and Systems*, ISCAS '09, March 2009.

[4] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge. Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits. *Proceedings of the IEEE, Special Issue on Ultra-Low Power Circuit Technology*, 98(2):253–266, February 2010.

[5] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger. Dark Silicon and the End of Multicore Scaling. In *International Symposium on Computer Architecture*, ISCA '11, pages 365–376, 2011.

[6] W. Huang, K. Rajamani, M. R. Stan, and K. Skadron. Scaling with Design Constraints: Predicting the Future of Big Chips. *IEEE Micro*, 31(4):16–29, July 2011.

[7] E. Krimer, R. Pawlowski, M. Erez, and P. Chiang. Synctium: a Near-Threshold Stream Processor for Energy-Constrained Parallel Applications. *IEEE Computer Architecture Letters*, 9(1):21–24, January 2010.

[8] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi. McPAT: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures. In *International Symposium on Microarchitecture*, MICRO '09, pages 469–480, 2009.

[9] H. H. Najaf-abadi, N. K. Choudhary, and E. Rotenberg. Core-Selectability in Chip Multiprocessors. In *International Conference on Parallel Architectures and Compilation Techniques*, pages 113–122, 2009.

[10] Nanoscale Integration and Modeling (NIMO) Group. Predictive Technology Model (PTM). `http://ptm.asu.edu`.

[11] J. M. Rabaey, A. Chandrakasan, and B. Nikolić. *Digital Integrated Circuits: A Design Perspective*. Prentice Hall, 2^{nd} edition, 2003.

[12] T. Sakurai and A. Newton. Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas. *IEEE Journal of Solid-State Circuits*, 25(2):584–594, Apr. 1990.

[13] G. Venkatesh, J. Sampson, N. Goulding, S. Garcia, V. Bryksin, J. Lugo-Martinez, S. Swanson, and M. B. Taylor. Conservation Cores: Reducing the Energy of Mature Computations. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '10, pages 205–218, 2010.